Tijdschrift voor Medisch Onderwijs, februari 2009 | Vol. 28, nr. 1, p. 13-21

Tentamens met de computer: een vergelijking van meerkeuzevragen en alternatieve vraagvormen

S. Draaijer, G.C. van den Bos.

Samenvatting

In het medisch onderwijs wordt steeds meer gebruik gemaakt van beeldschermtoetsing en alternatieve vraagvormen. In een speciaal geprepareerd beeldschermtentamen werden traditionele meerkeuzevragen en alternatieve vraagvormen aan studenten voorgelegd om deze voor wat betreft scores en slaagpercentages met elkaar te kunnen vergelijken. De meerkeuzevragen dienden tevens als referentie. Voor het tentamen werden drie alternatieve vraagvormen gebruikt: Drag-and-dropvragen, Multiple Responsevragen en Matchingvragen. De slaaggrens werd bepaald volgens het model waarbij de helft van het aantal punten gescoord moet worden nadat een correctie voor het scoren op basis van de raadkans is toegepast. Deze methode wordt in het hoger onderwijs op grote schaal gebruikt. De resultaten laten zien dat de alternatieve vraagvormen vele mogelijkheden bieden en scores opleveren in dezelfde ordegrootte als meerkeuzevragen. De alternatieve vraagvormen resulteren echter wel in verschillende slaagpercentages. (Draaijer S, Bos GC van den. Tentamens met de computer: een vergelijking van meerkeuzevragen en alternatieve vraagvormen. Tijdschrift voor Medisch Onderwijs 2009;28(1):13-21.)

Inleiding

Ondanks een enorme toename van het gebruik van digitale leeromgevingen, worden in het geneeskunde onderwijs nog maar weinig tentamens afgenomen via het beeldscherm. Dat heeft ten eerste een logistieke oorzaak: de meeste instellingen beschikken niet over zalen met voldoende computers voor alle examinandi. Een tweede reden is waarschijnlijk de onbekendheid met de mogelijkheden van moderne toetsprogramma's waarmee andere dan de klassieke meerkeuzevragen gegenereerd kunnen worden, zoals vragen waarbij de student begrippen of tekens moet verslepen naar locaties in een gepresenteerd figuur. Uit de literatuur is betrekkelijk weinig bekend over het gebruik van dergelijke alternatieve vraagvormen.[1-2] Richtlijnen voor het ontwikkelen van dergelijke vragen, in dit artikel 'alternatieve vraagvormen' genoemd, worden beschreven door bijvoorbeeld Draaijer en Hartog.[3-4] Met betrekking tot alle vraagvormen geldt echter dat het bij de student opgeroepen denkproces meer wordt bepaald door de stimulus van de vraag (wat er gevraagd wordt) dan door de responsvorm zoals de meerkeuze-, juist/onjuist-vraag of essay.[5] Wel kan een attractieve vormgeving en interactie bijdragen aan het verhogen van de motivatie van de lerende.[6-7] Verder is bekend dat polytoom gescoorde vragen (waarbij de student 0, 1, 2 etc. punten per vraag kan scoren) ten opzichte van dichotoom gescoorde vragen (waarbij de student slechts 0 of 1 punt kan scoren), betere vraag-toetscorrelaties hebben, maar meer tijd kosten om te beantwoorden.[8] Verder blijkt dat het bepalen van goede scoringsvoorschriften voor alternatieve vraagvormen niet eenvoudig is.[9-10]

Sinds twee jaar worden in het VUmc, als voorbereiding op de invoering van

beeldschermtoetsing, de herkansingen van de cursussen Bioregulatie, Hart en bloedsomloop en Nier en milieu-interieur – alle uit het tweede studiejaar van het aflopende Curriculum '91 – afgenomen met behulp van het programma Questionmark Perception (QMP). Dat programma ondersteunt een groot aantal vraagvormen via het beeldscherm.

In dit onderzoek hebben we onderzocht of door alternatieve vraagvormen het slagingspercentage verandert, en zo ja, hoe.

Methoden

Wij voerden ons experiment uit met de tweede herkansing van het derdejaars tentamen Voeding en Spijsvertering, waarin de vakgebieden Celbiologie (3), Medische Chemie (10), Fysiologie (12), Pathologie (10), Kindergeneeskunde (10), Heelkunde (10) en Maag/darm/leverziekten (23) een aandeel hadden (de getallen tussen haakjes geven het aantal vragen aan). Het tentamen bestond uit 56 vierkeuzevragen en 22 alternatieve vragen. De vierkeuzevragen en alternatieve vragen waren evenredig per onderwerp verdeeld. Aan het tentamen namen 70 studenten deel. Op grond van eerdere ervaringen met beeldschermtoetsing en de mogelijkheid om desnoods alleen van de 56 vierkeuzevragen gebruik te maken voor de becijfering, werd dit experiment verantwoord geacht.

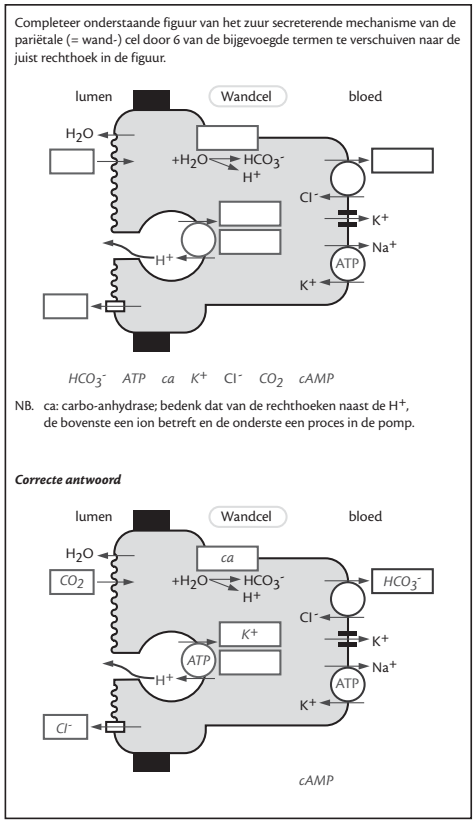
In het voorgaande reguliere tentamen, en de daaropvolgende eerste herkansing waren 75 vierkeuzevragen opgenomen. In het huidige tentamen (de tweede herkansing) werden de uitkomsten van de 56 vierkeuzevragen gebruikt als referentie: op basis van die vragen kon bepaald worden of alternatieve vraagvormen leiden tot hogere scores en hoe dit invloed heeft op het slaagpercentage. Naast de vierkeuzevragen kozen wij als alternatieve vragen:

- *Drag-and-dropvragen* (DrandDr); hierbij moet een aantal begrippen of symbolen versleept worden naar rechthoeken in een gegeven figuur, of naar een flowdiagram (aantal te verslepen termen één maal of meer groter dan het aantal rechthoeken of het aantal open plaatsen in het diagram om de onderlinge afhankelijkheid van de antwoordopties te verkleinen).
- *Matchingvragen*; hierbij moet een aantal begrippen uit één kolom op de juiste wijze worden gecombineerd met de begrippen in een tweede kolom (aantal begrippen in de tweede kolom één maal of meer groter dan in de eerste kolom om de onderlinge afhankelijkheid van de antwoordopties te verkleinen).
- *Multiple Responsevragen (MR);* hierbij is meer dan één keuze juist. Deze vragen worden ook vaak 'meer-uit-meer'-vragen genoemd.
  Het gekozen scoringsmodel is zo dat:
  - er per correct gekozen alternatief 1 punt wordt gegeven;
  - er per gekozen afleider 0 punten worden gegeven;
  - er vooraf bekend wordt gemaakt aan de student hoeveel juiste keuzes er zijn. Dit laatste wordt gedaan om de onzekerheid omtrent de wijze van beantwoorden te verkleinen.

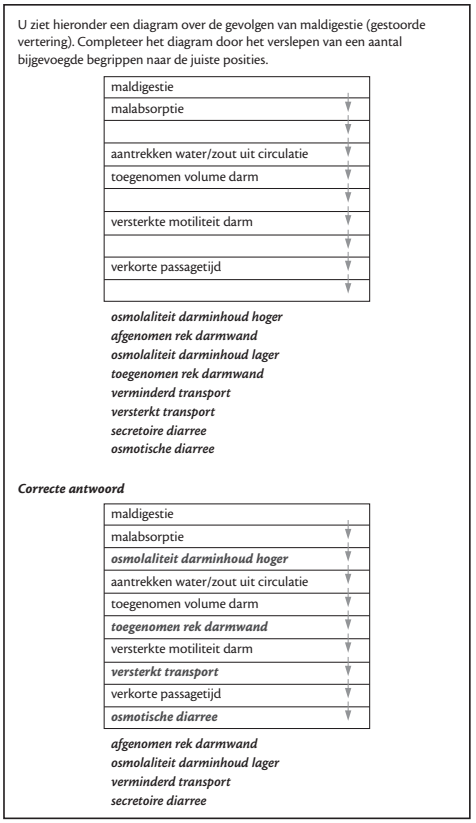
De figuren 1, 2, 3 en 4 zijn voorbeelden van respectievelijk een DrandDr- (Figuur 1 en 2), een Match- (Figuur 3) en een MR-vraag (Figuur 4).

Vier weken vóór de tentamendatum stond de studenten een oefententamen (inclusief oefenvoorbeelden van de alternatieve vragen) ter beschikking met beredeneerde en naar de studiestof verwijzende antwoorden.

De vierkeuzevragen werden gemaakt door de vakdocenten; voor de alternatieve vragen gaven deze docenten ook aan wat zij wilden vragen, waarna hun voorstellen door een QMP-deskundige werden omge-



Figuur 1. *Een DrandDr-vraag waarbij zes begrippen naar een basisfiguur moeten worden versleept (en er 1 afleider is: cAMP).*



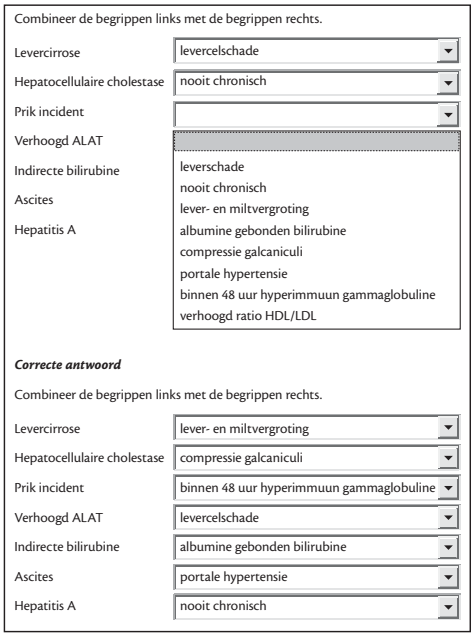
Figuur 2. *Een DrandDR-vraag waarbij een redenering met vier begrippen moet worden gecomplementeerd (er zijn dus ook vier afleiders).*

zet in het bedoelde QMP-format. De eindredactie van alle vragen, ook van de alternatieve vragen, lag bij de blokvoorzitter. De alternatieve vraagtypen werden polytoom gescoord: elke goede keuze in een vraag (bijvoorbeeld het selecteren van een goed alternatief uit een MR-vraag), leverde de student een punt op. Voor dit scoringsmodel is gekozen omdat daarbij elke goede keuze beloond wordt, wat door de studenten als meest redelijk wordt ervaren.

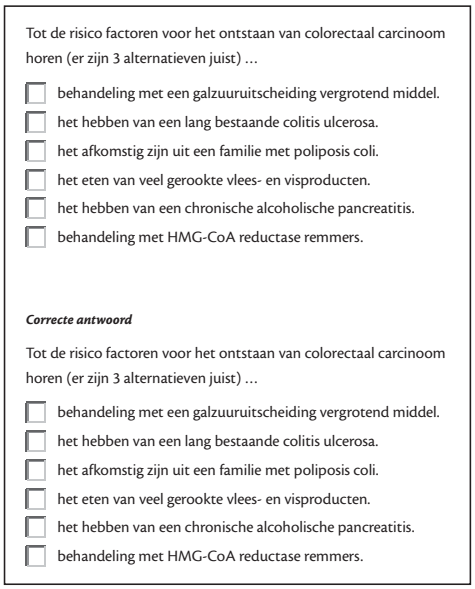
De score-cijfertransformatie gebeurde op basis van een lineair verband met verdiscontering van de raadscore. De raadscore wordt daarbij gedefinieerd als de verwachte score bij het volledig random beantwoorden van vragen.

In principe is die raadscore (q_{raad}) gelijk aan de optelsom van de kans op i=0, 1, 2 etc. punten ($p(q_i)$), maal het betreffende aantal punten (q_i). In formulevorm: $q_{raad}=\Sigma(p(q_i)*q_i)$.

De raadscore van een vierkeuze-multiple choice vraag is daarmee gelijk aan $q_{raad\_4mc}=p(0)*0+p(1)*1=0{,}75*0+0{,}25*1=0{,}25=25\%$. Bij toetsvragen die polytoom gescoord worden is het iets moeilijker een dergelijke raadscore te berekenen. Voor bijvoorbeeld een MR-vraag met vijf opties, waarbij drie opties correct zijn



Figuur 3. *Een Match-vraag waarbij acht begrippen in de rechterkolom, ieder met bijbehorend begrip uit de linkerkolom gecombineerd moeten worden (en er is 1 afleider: verhoogde ratio HDL/LDL).*

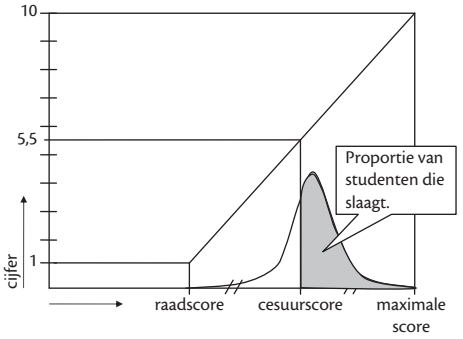


Figuur 4. *Een MR-vraag waarbij de drie juiste alternatieven aangevinkt moeten worden (op beeldscherm is de volgorde van alternatieven gerandomiseerd).*

volgens het scoringsmodel zoals gegeven op blz. 5, is deze kans $q_{raad\_5\text{-}3MR} = p(0) * 0 + p(1) * 1 + p(2) * 2 + p(3) * 3 = 0*0+4/10*1+5/10*2+1/10*3=1{,}70$ punten.

Voor Matchingvragen (waarbij elke te kiezen optie één maal gekozen mag worden) geldt een relatief eenvoudige berekening waarbij geldt dat de kansscore gelijk is aan de kans per optie at random gekozen te worden maal het aantal vragen: q_r=aantal_vragen*(1/aantal opties). Voor bijvoorbeeld een Matchingvraag met vijf vragen en zes opties is de kans $q_{raad\ 5\text{-}6\_Match}=5*1/6=5/6$.

Het geven van cijfers aan studenten vond plaats op basis van de behaalde score, verdisconteerd met de raadscore. De methode is grafisch weergegeven in Figuur 5. Deze methode wordt veel toegepast in het hoger onderwijs en is efficiënt.[11] In het kort komt de methode er op neer dat de score waarbij studenten slagen, halverwege de raadscore en de maximale score wordt gekozen (vaak wordt ook de grens op 55% of 60% gesteld). Het is belangrijk om daarbij aan te geven dat een kleine aanpassing van de cesuurscore een grote invloed heeft op het percentage geslaagde studenten. Dit wordt veroorzaakt doordat in veel toetsen de grootste groep studenten een score heeft die zich rondom die cesuurscore bevindt. Zo kan bijvoorbeeld een verhoging van de cesuurscore met 1% leiden tot mogelijk 6% minder geslaagden op een toets. In Figuur 5 wordt dit aangegeven door de verticale lijn bij de cesuurscore die het gebied afsluit van de proportie van de studenten die slaagt.



Figuur 5. *Het verband tussen de score op een toets en het bijbehorende cijfer, de cesuurscore en het slaagpercentage.*

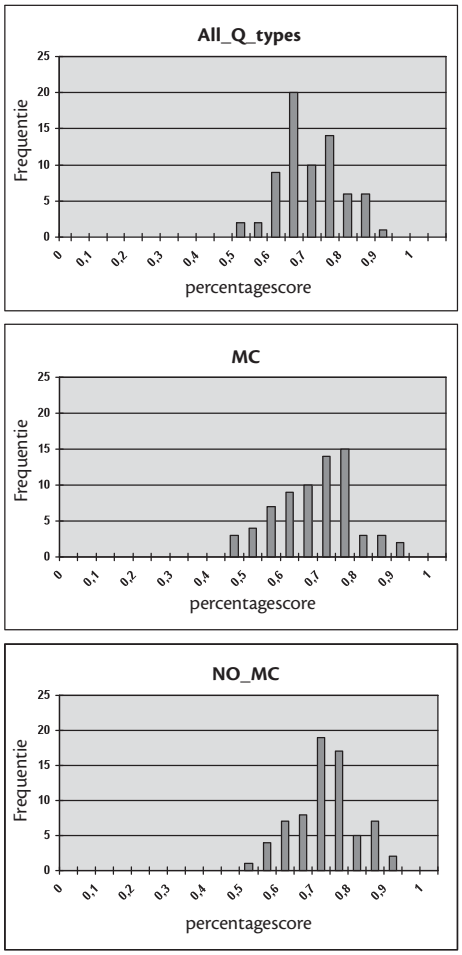
De figuur toont dat het laagste cijfer (bijv. een 1,0) wordt toegekend aan studenten die de raadscore behalen, en het cijfer 10,0 aan studenten die alle vragen correct beantwoorden (maximale score). De cesuur (bijv. het cijfer 5,5) ligt halverwege de raadscore en de maximale score. De normaalcurve geeft de gebruikelijke spreiding van de scores aan over de populatie van de studenten. Afhankelijk van deze spreiding is er een bepaald percentage studenten dat zakt cq. slaagt.

Resultaten

Na een vraaganalyse bleken vier vierkeuzevragen ongeschikt, vanwege de combinatie van een zeer lage score met een lage of negatieve vraag-toetscorrelatie (R_{it}) wat meestal duidt op een vraag die niet duidelijk is geformuleerd. Ook werden twee MR-vragen uit de analyse verwijderd omdat in de stam van die twee vragen niet het aantal correcte alternatieven bleek te zijn aangegeven terwijl dat bij alle andere MR-vragen wel het geval was. Het opnemen van deze twee MR-vragen in de analyse zou kunnen leiden tot uitspraken over twee ongelijke vraagvormen.

De scoreverdeling op de toets is weergegeven in de histogrammen van Figuur 6. De scoreverdeling toont een normaalverdeling met verschoven top. Deze is gebruikelijk voor tentamens.

De betrouwbaarheid (Cronbach Alpha) van het tentamen als geheel was 0.79. Dat is hoger dan die van de reguliere (papieren) herkansingen (0.72), en voor summatieve toetsen een acceptabele waarde. De betrouwbaarheid van de toets op basis van *alleen* de meerkeuzevragen was 0.70 en op basis van *alleen* de alternatieve vragen 0.65. Deze drie waarden bevestigen dat de toets goed discrimineert tussen de studenten.



Figuur 6. *Histogrammen van de scoreverdeling op de toets voor alle vragen gezamenlijk (All_Q_types), alleen de vierkeuzevragen (MC) en alternatieve vragen (NO_MC).*

Nadat de studenten hun antwoorden hadden 'ingeleverd' vroegen wij hen individueel naar hun ervaringen. Zij gaven aan dat zij het tentamen 'leuk en uitdagend' hadden gevonden, onder meer omdat zij vonden dat ze gedwongen werden om beter na te denken, vooral bij de DrandDr- en de Matchingvragen. Dit commentaar is in overeenstemming met de verwachtingen.

De gemiddelde score van de 70 studenten op het tentamen, uitgedrukt in een percentagescore, was 66%. Als deze score wordt vertaald naar een situatie waarbij bijvoorbeeld alleen het gebruik van dichotoom gescoorde meerkeuzevragen zou zijn toegepast, betekent dit dat de studenten gemiddeld 66% van de vragen goed hebben beantwoord. Indien deze score vertaald wordt naar polytome vragen, wil dit zeggen dat studenten gemiddeld 66% van het totaal aantal te behalen punten hebben gescoord.

Op basis van de score-cijfertransformatie als aangegeven in Figuur 5, zou 67% van de studenten slagen (dat het cijfer van de gemiddelde score en het slaagpercentage bijna gelijk zijn is in dit geval toevallig – zie de toelichting bij Figuur 5). In Tabel 1 zijn de aantallen vragen en hun karakteristieken weergegeven.

Uit Tabel 1 blijkt dat de *gemiddelde score* op de vragen uiteenloopt. Het gemiddelde tussen de score op de MC-vragen en de alternatieve vraagvormen verschilt 4% en dit verschil is significant ($t(69)=3.89$; $p=.000$.). De Match-vragen scoren bijna net zo hoog als de meerkeuzevragen; de MR- en de DrandDr-vragen scoren gemiddeld hoger. De gemiddelde vraag-toetscorrelatiewaarden van de alternatieve vragen zijn, zoals verwacht, ook hoger, maar de waarde van de MR-vraag wijkt nauwelijks af van de MC-vragen. Dat laatste is een indicator dat het meetvermogen van de MR-vragen niet optimaal is.

Tabel 1 *Overzicht van het tentamen en het aantal verschillende vragen met hun diverse bijbehorende karakteristieken.*

	Uitsluitend meerkeuze vragen[a]	Uitsluitend alternatieve vraagvormen[b]	MR	Match	DrandDr	Alle vragen
Aantal vragen (n)	52	20	6	6	8	72
Maximum score (punten)	52	92	18	41	33	144
Gemiddelde procentuele score op de vragen	65%	69%	71%	64%	72%	66%
Cronbach Alpha	0,70	0,65				0,79
Gemiddelde van vraag-toetscorrelatie [a]	0,21	0,34	0,29	0,44	0,30	0,25
Raadscore (punten) en bijbehorend raadpercentage	13 / 25%	22,8 / 25%	8,8 / 49%	5,2 / 13%	8,8 / 27%	38,8 / 26%
Cesuurscore (punten) [b]	32,5	57,4	13,4	22,6	20,9	89,9
Slaagpercentage van de studenten	61%	77%	34%	79%	81%	67%

[a] *Gemiddelde item-testcorrelatie (correlatie tussen de score op een vraag en de toetsuitslag) op een schaal van -1 tot +1; hoe hoger de waarde des te meer de score op de betreffende vraag correleert met de eindscore op de toets. In het algemeen dient deze waarde hoger dan 0,2 te zijn.*

[b] *De cesuurscore op basis van het lineaire scoringsmodel uitgaande van de raadscore en cesuur op 50%.*

In Tabel 1 is op basis van de scores en de score-cijfertransformatie aangegeven welk deel van de studenten zou slagen. Volgens deze gegevens zou op basis van *alleen* de vierkeuzevragen, 61% van de studenten slagen. Op basis van de alternatieve vraagvormen zou 77% van de studenten slagen. Dit is significant: Chi^2 (1)= 0.02; p=.006.). Op basis van alleen de MR-vragen zou 34% slagen, terwijl dat voor de Match-vragen 79% en de DrandDr-vragen 81% zou zijn.

Gezien het relatief grote aantal punten dat behaald kon worden bij de Match- en de DrandDr-vragen, is het uiteindelijke slaagpercentage ten opzichte van alleen de MC-vragen hoger. De invloed van het lage slagingspercentage op basis van alleen de MR-vragen is, vanwege het relatief lage aantal te behalen punten voor de MR-vragen, klein.

Discussie

We hebben een experiment uitgevoerd in een reële tentamensituatie. Studenten hebben een computertoets gedaan waarin ze antwoorden moesten geven op een mix van meerkeuzevragen en alternatieve vraagvormen. Doordat we kunnen aannemen dat de beheersingsgraad van de stof door de studenten gelijk is, kunnen we uitspraken doen over de verschillen in scores en slaagpercentages voor de verschillende vraagvormen.

Methodische beperkingen

Uit logistieke overwegingen (gelimiteerd aantal beeldschermen en ruimtes) hebben wij voor een herkansingstentamen gekozen. Het gebruik van groepen herkansers voor onderwijskundige experimenten is aanvechtbaar: de groepen zijn voorgeselecteerd vanwege het falen bij een eerste tentamenpoging en het aantal studenten is meestal klein. In ons geval moet ook nog in de overwegingen betrokken worden dat deze herkansers onder de druk van een ten einde lopend curriculum werkten. De scores van de studenten zijn echter volgens een normaalverdeling gespreid. De combinatie van daadwerkelijke kennisspreiding en normaalverdeling geeft aan dat de toets voldeed aan voorwaarden voor klassieke toetsanalyse. Aan de voorwaarden – uitspraken te doen over het tentamen als geheel – is dus voldaan.

Helaas is het niet mogelijk voor wat betreft de scores en de percentages geslaagden *algemene* conclusies te trekken voor de verschillende alternatieve vraagvormen. In het experiment zijn van elk type daarvoor te weinig vragen aanwezig (n=6, 8); conclusies in dit artikel zijn dan ook vooral illustratief bedoeld. Ze vormen vooral een aanzet tot discussie en verder onderzoek.

Overwegingen met betrekking tot de percentagescore en cesuurbepaling

Voordat wij tot het gebruik van alternatieve vragen overgingen, verwachtten wij dat dit zou leiden tot een lager slagingspercentage. Wellicht omdat zij meer gericht waren op inzicht dan op feiten. Het slagingspercentage daalde echter niet.

Als gekeken wordt naar de *gemiddelde score* op de vragen, blijken MR-vragen en DrandDr- vragen hoger te scoren dan vierkeuzevragen. Op deze twee vraagtypen kiezen studenten blijkbaar snel de juiste opties en scoren snel punten. De gemiddelde score op de Match-vragen is bijna gelijk aan die van de vierkeuzevragen.

Wordt er gekeken naar het *slagingspercentage*, dan zien wij een ander beeld. Toepassing van de score-cijfertransformatie op basis van de raadscore leidt bij MR-vragen tot een laag, en bij Match- en DrandDr-vragen tot een hoog slagingspercentage. Bij de MR-vragen wordt dit veroorzaakt doordat hun raadscore in de buurt van 50% ligt zodat de cesuurscore

bij MR-vragen hoog is. Hierdoor slagen weinig studenten bij een gemiddelde score die vergelijkbaar (of slechts weinig hoger) is met die van vierkeuzevragen. Bij de Match-vragen slagen relatief veel studenten omdat de raadscore en de bijbehorende cesuurscore juist vrij laag zijn. Bij de DrandDr-vragen is het beeld opnieuw anders: hier slagen meer studenten terwijl de raadscore vergelijkbaar of zelfs kleiner is dan die van meerkeuzevragen. Voor de DrandDr-vragen scoren de studenten blijkbaar daadwerkelijk gemakkelijker punten.

Voor de MR-vragen uit het experiment kunnen we zeggen dat de keuze voor het scoringsmodel voor deze vragen (1 punt per goed gekozen alternatief en geen aftrek van punten voor een gekozen afleider) niet leidt tot een betrouwbare meting. Door de hoge raadkans wordt het meetgebied van deze vragen klein en discrimineren de vragen niet heel goed (gemiddelde R_{it} waarde is niet hoger dan vierkeuzevragen). Het werken met aftrek van punten per gekozen afleider – en bijvoorbeeld een minimale score van 0 punten – kan leiden tot een beter scoringsmodel. Hierdoor daalt de raadscore sterk.

Voor zowel de Match- als de DrandDr-vragen in het experiment zouden we kunnen zeggen dat de studenten hun partiële kennis goed kunnen laten zien en daar ook voor beloond worden. Ze scoren waarschijnlijk relatief gemakkelijk punten door het goed kiezen van de 'gemakkelijke' onderdelen van deze vragen. De 'moeilijkere' onderdelen zorgen er echter voor dat de vragen toch goed discrimineren. Door de interne afhankelijkheid van de afleiders in dergelijke vragen (ondanks het invoegen van afleiders bij deze vragen) leidt de statistische raadscore tot een hoger slaagpercentage dan bij toepassing van alleen vierkeuzevragen. Bij de cesuurbepaling zou daar rekening mee kunnen worden gehouden.

Overwegingen voor ontwikkeling van alternatieve vragen of wel beeldschermvragen

De validiteit van een toets wordt bepaald door de mate waarin de toets meet wat gemeten dient te worden. Het doel van het maken van alternatieve vragen voor tentamens is niet dat op basis daarvan evenveel studenten slagen als bij het alleen toepassen van MC-vragen. Het doel moet zijn studenten te bevragen op een aantrekkelijke manier en op een manier die recht doet aan de stof. Een zeer groot deel van de geneeskundige diagnostiek berust op beeldvormende technieken. Studenten moeten daartoe omgaan met foto's en afbeeldingen, waarop zij afwijkingen moeten leren herkennen en aanwijzen: DrandDr-technieken zijn ideaal om dergelijke vaardigheden te onderwijzen en te toetsen. Veel inzicht in de geneeskunde berust bovendien op concepten waarvan kennis inzichtelijk kan worden gemaakt en getoetst door het verslepen van begrippen naar 'lacunes in een diagram of flowchart'. Ook hier dus een vorm van de DrandDr-technieken. Een bijkomend argument voor het toepassen van dergelijke vragen is dat beelden vaak eenduidiger zijn dan tekst, waardoor de invloed van de taalbeheersing van studenten op de meting wordt verminderd.

Onze bevindingen laten zien dat alternatieve vraagvormen niet onder doen voor klassieke vragen bij het maken van toetsen. Zij maken het mogelijk om leerstof interessanter te bevragen. De vragen maken goed onderscheid in de mate waarin de studenten de stof beheersen. Voordat het geneeskunde onderwijs echter op grote schaal van dergelijke alternatieve vragen gebruik zal kunnen maken, is meer inzicht nodig in de specifieke eigenschappen van deze vragen en moeten docenten een gevoel ontwikkelen voor de 'moeilijkheid' van dergelijke vragen. Proe-

ven met gelijke aantallen meerkeuzevragen en alternatieve vraagvormen en grotere aantallen studenten zijn daar in het bijzonder voor nodig.

Online voorbeelden van digitale vragen zijn te bekijken via: *https://www.surfgroepen.nl/sites/flextoets/NVMOartikel/Home.aspx* Gebruik hiervoor Internet Explorer aangezien andere browsers de toetsvragen niet goed weergeven.

Dankwoord

Wij danken drs. C.J.L.H. Camps, dr. A.J. Greven, dr. W. van de Laarse, dr. R.J.P. Musters, drs. C. Reumer en drs. M.I. Schade voor hun adviezen en het kritisch lezen van het manuscript.

Literatuur

1. Haladyna TM. Developing and Validating Multiple-Choice Test Items. Third Edition ed. London: Lawrence Erlbaum Associates; 2004.
2. Parshall CG, Spray JA, Kalohn JC, Davey T. Practical considerations in computer-based testing. New York: Springer-Verlag; 2002.
3. Draaijer S, Hartog R. Design Patterns for digital item types in Higher Education. E-Journal of Instructional Science and Technology 2007;10(1).
4. Draaijer S, Hartog R. Guidelines for the Design of Digital Closed Questions for Assessment and Learning in Higher Education. E-Journal of Instructional Science and Technology 2007:10(1).
5. Schuwirth LWT, van der Vleuten CPM. ABC of learning and teaching in medicine: Written assessment. BMJ 2003 March 22, 2003;326(7390):643-5.
6. Keller JM. Development and Use of the ARCS Model of Motivational Design. Enschede: Twente University of Technology; 1983. Report No.: IR 014 039.
7. Ricketts C, Wilks S, Crocker C. What factors affect student opinions of Computer-Assisted Assessment? 5th CAA Conference; 2001; Loughborough; 2001.
8. Jodoin MG. Measurement Efficiency of Innovative Item Formats in Computer-Based Testing. Journal of Educational Measurement 2003;40(1): 1-15.
9. Lampe T, Eggen T. Innovative Item Types in Computer Based Testing: Scoring of Multiple Response Items. Arnhem, The Netherlands: Citogroep; 2003.
10. Bull J, McKenna C. Blueprint for Computer-assisted Assessment: RoutledgeFalmer; 2001.
11. Dousma T, Horsten A, Brants J. Tentamineren. Derde druk ed. Groningen: Wolters-Noordhoff; 1997. [Testing. Groningen: Wolters-Noordhoff; 1997].

De auteurs:
S. Draaijer is onderwijskundig adviseur, verbonden aan het Onderwijscentrum van de VU Amsterdam.
G.C. van den Bos is werkzaam bij het Onderwijsinstituut en bij de afdeling Fysiologie van het VUmc.

Correspondentieadres:
Ir. S. Draaijer, Onderwijscentrum VU, Vrije Universiteit Amsterdam. E-mail: s.draaijer@ond.vu.nl

Belangenconflict: geen gemeld
Financiële ondersteuning: geen gemeld

Summary

The use of on-screen computer-based assessment in medical education is increasing. Students were given a computer-based test containing traditional multiple-choice questions and new question types and we compared the scores and pass rates of the different question types: drag-and-drop questions, multiple response questions and matching questions. The multiple choice questions served as base-line. The cut-scores were set according to the linear model for score-to-grade transformation adapted on the basis of the guess score. This method is common practice in higher education in the Netherlands. The results of the experiment show that the new question types reliably assess medical knowledge and that despite variation the scores are comparable to those of multiple choice questions. The new question types lead to considerable differences in passing rates, however. (Draaijer S, Bos GC van den. Computer-based assessment: a comparison between multiple-choice questions and new question types. Dutch Journal of Medical Education 2009;28(1):13-21.)